

Supplementary Information (SI)

1 Details of the environmental data processing and estimated kernel weights

Supplementary Figure S1 presents the weights estimated for each environmental data source. Table S1 presents the preprocessings that were

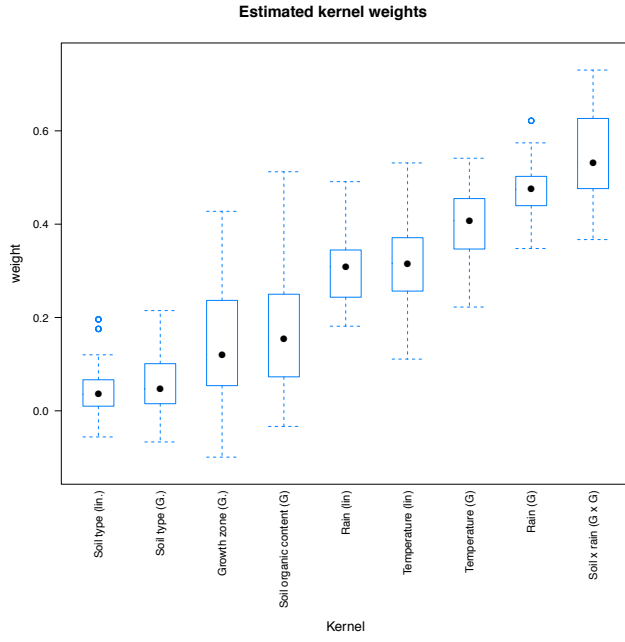


Fig. S1. Sensitivity analysis of the estimated normalized kernel weights.

applied to the different environmental covariates and the kernel functions used.

2 Details of the variational inference algorithm

For short-hand, the hyper-parameters in the model are denoted jointly by

$$\zeta = \{\alpha_j, \beta_j, \sigma_{g0}^2, \sigma_{e0}^2, \sigma_g^2, \sigma_e^2, \lambda_{g0}, \lambda_g, \lambda_e\},$$

and the parameters by

$$\Theta = \{\mathbf{a}_{g0}, A_g, A_e, H_g, H_e, \mathbf{g}^*, \mathbf{e}^*, \sigma_*^2\},$$

where $\sigma_*^2 = (\sigma_1^2, \dots, \sigma_{N_e}^2)$. In the following the dependence on ζ is omitted for clarity. We assume the factorized variational approximation

$$p(\Theta | K_g, K_e, Y) \approx q(\Theta) = q(\mathbf{a}_{g0})q(A_g)q(A_e)q(H_g)q(H_e)q(\mathbf{g}^*)q(\mathbf{e}^*)q(\sigma_*^2)$$

and define each factor in the ensemble just like its full conditional:

$$q(\mathbf{a}_{g0}) = \mathcal{N}(\mathbf{a}_{g0}; \mu(\mathbf{a}_{g0}), \Sigma(\mathbf{a}_{g0}))$$

$$q(A_g) = \prod_{r=1}^R \mathcal{N}(\mathbf{a}_r^g; \mu(\mathbf{a}_r^g), \Sigma(\mathbf{a}_r^g))$$

$$q(A_e) = \prod_{r=1}^R \mathcal{N}(\mathbf{a}_r^e; \mu(\mathbf{a}_r^e), \Sigma(\mathbf{a}_r^e))$$

$$q(H_g) = \prod_{i=1}^{N_g} \mathcal{N}(\mathbf{h}_i^g; \mu(\mathbf{h}_i^g), \Sigma(\mathbf{h}_i^g))$$

$$q(H_e) = \prod_{j=1}^{N_e} \mathcal{N}(\mathbf{h}_j^e; \mu(\mathbf{h}_j^e), \Sigma(\mathbf{h}_j^e))$$

$$q(\mathbf{g}^*) = \prod_{i=1}^{N_g} \mathcal{N}(g_i; \mu(g_i), \Sigma(g_i))$$

$$q(\mathbf{e}^*) = \prod_{j=1}^{N_e} \mathcal{N}(e_j; \mu(e_j), \Sigma(e_j))$$

$$q(\sigma_*^2) = \prod_{j=1}^{N_e} \mathcal{G}(\sigma_j^{-2}; \alpha(\sigma_j^{-2}), \beta(\sigma_j^{-2})).$$

The parameters in the factor distributions can be derived as by Gönen and Kaski (2014), and they are therefore omitted from here.

Initialisation of the variational algorithm. The parameter \mathbf{g}^* was initialised to the main genetic effects learnt by GBLUP, and \mathbf{e}^* was initialised to the average yields in the different environments. Parameters H_g and H_e were initialised by applying the regularized Singular Value Decomposition (SVD) implemented in R library `softImpute` to the yield matrix Y after regressing out the initialised main effects \mathbf{g}^* and \mathbf{e}^* . Parameters \mathbf{a}_{g0} , A_g and A_e were initialised to 0. Environment-specific residual variance parameters σ_*^2 were initialised to environment-specific sample variances.

3 Gains from modelling $G \times E$ for current target population of environments

Our results indicate targeted breeding could improve yields by dividing a single target population of environments (TPE) into several parts, but the same methodology could be used even when developing only 1 variety for a larger population of target environments as in traditional breeding. Traditional breeding makes the implicit assumption that varieties' observed yields $g \in 1, \dots, G$ in trial experiments in environments (location \times year) $e \in 1, \dots, E$, are representative of the yield in the TPE, in other words

$$p(\text{yield}_g | \text{TPE}) \approx \frac{1}{E} \sum_e p(\text{yield}_g | \text{environment}_e) \quad (1)$$

However, with geographic field use information and weather data widely available, this strong assumption can be replaced with an estimate for the yield in the TPE given the actual fields and their microclimates:

$$p(\text{yield}_g | \text{TPE}) \approx \sum_f^F P_f \times p(\text{yield}_g | f) \quad (2)$$

$$= \sum_f^F P_f \times \int_{\theta_f} p(\text{yield}_g | \theta_f) \times p(\theta_f) d\theta_f, \quad (3)$$

where $f \in 1, \dots, F$, are fields in the TPE used for cultivation of the new variety, θ_f are parameters (e.g. weather conditions) related to a certain

Variable (unit)	transformation	preprocessing parameters	missing value imputation	kernel transformation(s)
Soil content (% , $N_{\text{covs}} = 3$)	log transformation	z-normalization	(none)	linear and Gaussian
Soil organic content (% , $N_{\text{covs}} = 3$)	log transformation	z-normalization	(none)	Gaussian
daily rainfall (mm , $N_{\text{covs}} = 123$)	7-day moving average (6 previous days)	z-normalization with 3 rd order polynomial smoothing	0-imputation	linear and Gaussian
daily average temperature (C° , $N_{\text{covs}} = 123$)		z-normalization with 3 rd order polynomial smoothing of daily mean/scale parameters	0-imputation	linear and Gaussian
growth zone (1-4 , $N_{\text{covs}} = 1$)		z-normalization	(none)	Gaussian
genotype markers (SNPs , $N_{\text{covs}} = 5696$)		Minor allele frequency scaling for SNP A: $\frac{A - 2 \cdot \text{MAF}_A}{\sqrt{2 \cdot \text{MAF}_A \cdot (1 - \text{MAF}_A)}}$	mean imputation	linear kernel

Table S1. Preprocessings and kernel functions applied to covariates.

field f , $p(\theta_f)$ is the uncertainty related to these conditions, estimated from historical records, $p(\text{yield}_g | \theta_f)$ is the predictive distribution for the yield under conditions θ_f , obtained from the model, and P_f is the proportion of the total volume cultivated in field f .

4 Details of the cross validation scheme

The data consists of two generations of lines: a parental and progeny generation. Data is available from several years and locations.

Algorithm 1 presents the pseudo code for the 3D nested cross validation used to evaluate performance in new location, on new years for new genotypes.

Table S2 presents a detailed comparison of the proposed cross-validation setup to earlier works.

5 Climatic variation between the trial locations

Statistics about the weather conditions on trial locations on different months and years is presented in Figure S2.

6 Details of the cross validation split

Table S3 presents statistics about the cross validation splits.

7 Details of the sensitivity analysis

In the sensitivity analysis, the effect of the removal of training environments (removing all data from one further location and year from the training set) is obtained by evaluating test set performance with the

models trained during cross validation after omitting different validation folds (line 22 in pseudo code): test set performance was evaluated with each of these models to measure the sensitivity of model performance to the decomposition of the training set.

The number of test folds and validation folds corresponding to each location-year combination varies (see Table S3) due to availability of observations in the data set. The total number of validation folds for the 41 test folds is 343. Information about training and test set sizes and the

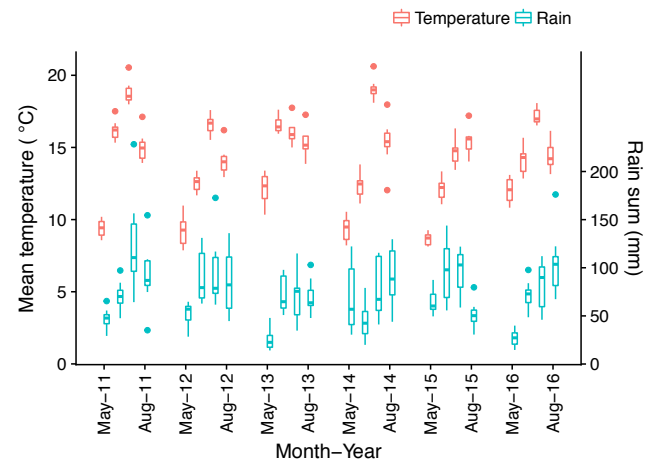


Fig. S2. Climatic variation among the different trial locations on different months and years. The dashed lines denote the mean \pm standard deviation.

number of validation folds for different cross validation splits is given in Supplementary Table S3.

Algorithm 1 3D Nested Cross Validation for New Year, New Location and New Genotype -performance estimation

```

1:  $L$ : set of all locations
2:  $Y$ : set of all years
3:  $G_{parental}$ : set of all lines in the parental population
4:  $G_{progeny}$ : set of all lines in the progeny population
5:  $\Theta$ : set of model hyperparameter combinations

6: Initialise training set (all data), validation set (empty set) and test set (empty set)

7: Assign test set:

8: for test location in  $L$  do
9:   for test year in  $Y$  do
10:    for set of test genotypes in  $G_{progeny}$  do
11:      Assign all observations from { test location, test year and set of test genotypes } as the test set.
12:      Omit all other data from the test year, location or lines from the training set (not only combination, each condition individually)
13:      Assign validation set:
14:      for validation location in  $\{L \setminus \text{test location}\}$  do
15:        for validation year in  $\{Y \setminus \text{test year}\}$  do
16:          for set of validation genotypes in  $G_{progeny} \setminus \text{test genotypes}$  do
17:            Assign all observations from { validation location, validation year and set of validation genotypes } as the validation set.
18:            Omit all other data from the validation year, location or lines from the training set (not only combination, each condition individually)
19:            Omit all other data  $G_{progeny}$  from the training set
20:            for model hyperparameters  $\theta$  in  $\Theta$  do
21:              Train model on training set
22:              Compute performance on validation set
23:            end for
24:          end for
25:        end for
26:      end for
27:      Select hyperparameter combination by averaging over the different validation sets.
28:      Assign training set without omitting any validation set, however, omit all data from  $G_{progeny}$ 
29:      Train model on the training set
30:      Compute performance on the test set
31:    end for
32:  end for
33: end for

34: Estimate generalisation performance by averaging over the different test sets.

```

Publication	New environment	New genotypes
Burgueño <i>et al.</i> (2012) (CV1/CV2)	CV1/CV2: test locations and years are present in the location-year combinations in the training data	new lines in CV1: not restricted to the offspring generation. In CV2 the test lines have phenotype observations
Heslot <i>et al.</i> (2014)	Random split, balanced wrt years and locs → years and locations not new	only 544/2195 genotypes have no phenotype observations, test set not restricted to the offspring generation
Albrecht <i>et al.</i> (2014)	the year-location combination is new but the test locations and years are present in other location-year combinations in the training data	genotypes are new and from the offspring
Malosetti <i>et al.</i> (2016)	time-structured DTD: 2/6 test locations new according to strict criteria; physically structured DTD: none of the environments are strictly new (as the year is not new)	all genotypes within the same family, not from the next generation.
Saint Pierre <i>et al.</i> (2016) (leave-one-side-out)	location new but year part of the training set	test lines have phenotype observations
Jarquín <i>et al.</i> (2017)	CV00: new location-year combination but the location and year are present in other combinations. CV0: new location or year	CV00: new genotypes but not from the offspring generation. CV0: phenotype observations available for the genotypes

Table S2. Comparison of the proposed in silico setup to the existing setups.

environment	n phenotypes final training set	n validation folds	n phenotypes nested CV mean \pm sd	n phenotypes validation mean \pm sd	yield mean \pm sd	n phenotypes test set
1 Loc B, 2011	5676	10	4197 \pm 557	113 \pm 46	6685 \pm 585	59
2 Loc A, 2011	7049	10	4778 \pm 998	113 \pm 49	5509 \pm 741	59
3 Loc G, 2011	6615	10	4519 \pm 953	121 \pm 57	6054 \pm 620	58
4 Loc B, 2013	5205	9	4028 \pm 492	95 \pm 58	5061 \pm 565	182
5 Loc A, 2012	7294	10	4977 \pm 1084	94 \pm 36	5799 \pm 829	106
6 Loc G, 2012	6709	10	4529 \pm 1042	94 \pm 36	5186 \pm 745	106
7 Loc D, 2012	8406	10	5994 \pm 1072	96 \pm 38	3617 \pm 667	105
8 Loc E, 2013	6990	10	4972 \pm 798	103 \pm 71	7178 \pm 589	91
9 Loc B, 2012	5613	10	4168 \pm 557	97 \pm 39	4735 \pm 768	106
10 Loc G, 2013	5902	10	4086 \pm 823	105 \pm 67	5223 \pm 1025	91
11 Loc B, 2012	5613	8	4126 \pm 559	80 \pm 21	5414 \pm 768	260
12 Loc A, 2012	7294	10	4854 \pm 1038	90 \pm 36	5760 \pm 780	243
13 Loc E, 2013	6990	10	4972 \pm 746	84 \pm 23	6948 \pm 752	153
14 Loc G, 2013	5902	10	4102 \pm 818	86 \pm 23	5557 \pm 811	152
15 Loc C, 2014	8391	10	5786 \pm 1018	89 \pm 35	3978 \pm 481	79
16 Loc B, 2013	5205	9	3891 \pm 517	86 \pm 23	5932 \pm 491	153
17 Loc A, 2013	6141	10	4293 \pm 828	86 \pm 23	7096 \pm 766	153
18 Loc E, 2014	8231	10	5590 \pm 953	89 \pm 36	4858 \pm 556	79
19 Loc B, 2014	5769	9	4022 \pm 696	83 \pm 21	4408 \pm 398	79
20 Loc B, 2014	5769	10	3816 \pm 631	120 \pm 53	4130 \pm 637	106
21 Loc C, 2014	8391	10	5491 \pm 888	130 \pm 61	4886 \pm 784	106
22 Loc E, 2014	8231	10	5352 \pm 795	131 \pm 64	5378 \pm 852	105
23 Loc H, 2015	7450	10	4792 \pm 933	122 \pm 59	4411 \pm 379	64
24 Loc F, 2015	8205	10	5524 \pm 935	122 \pm 59	7287 \pm 694	64
25 Loc B, 2015	5610	10	3778 \pm 682	113 \pm 51	5699 \pm 903	64
26 Loc B, 2013	5731	6	4758 \pm 371	68 \pm 19	5167 \pm 1095	488
27 Loc G, 2013	6703	9	4805 \pm 1029	68 \pm 20	5544 \pm 892	244
28 Loc E, 2013	7996	8	6036 \pm 1017	65 \pm 19	7175 \pm 782	244
29 Loc C, 2014	9393	5	7190 \pm 1273	55 \pm 12	5232 \pm 499	120
30 Loc F, 2015	9409	3	7546 \pm 1493	91 \pm 0	7053 \pm 811	39
31 Loc E, 2014	9172	6	7034 \pm 1275	56 \pm 11	5600 \pm 459	120
32 Loc B, 2015	6432	2	5642 \pm 276	91 \pm 0	7463 \pm 489	39
33 Loc B, 2014	6504	6	5212 \pm 891	119 \pm 97	3571 \pm 591	91
34 Loc E, 2014	9172	8	6748 \pm 1393	133 \pm 158	5485 \pm 791	91
35 Loc C, 2014	9393	8	6838 \pm 1422	156 \pm 160	4494 \pm 606	91
36 Loc H, 2015	8652	5	6170 \pm 1526	243 \pm 150	4013 \pm 554	42
37 Loc B, 2015	6432	4	4881 \pm 1037	182 \pm 72	6537 \pm 659	42
38 Loc B, 2015	6432	6	5135 \pm 903	152 \pm 73	7829 \pm 596	64
39 Loc F, 2015	9409	8	7159 \pm 1439	186 \pm 138	6167 \pm 924	64
40 Loc H, 2015	8652	8	6402 \pm 1439	186 \pm 138	7224 \pm 671	63
41 Loc C, 2015	9111	6	6531 \pm 1445	213 \pm 152	5356 \pm 1024	60

Table S3. Statistics about the different training, validation and test splits used in the cross validation.